基于 GRU 和注意力机制的远程监督关系抽取 *

黄兆玮，常　亮，宾辰忠†，孙彦鹏，孙　磊

(桂林电子科技大学 广西可信软件重点实验室，广西 桂林 541004)

摘　要：随着深度学习的发展，越来越多的深度学习模型被运用到了关系提取的任务中，但是传统的深度学习模型无法解决长距离依赖问题；同时，远程监督将会不可避免地产生错误标签，针对这两个问题，提出一种基于 GRU（gated recurrent unit）和注意力机制的远程监督关系抽取方法，首先通过使用 GRU 神经网络来提取文本特征，解决长距离依赖问题；接着在实体对上构建句子级的注意力机制，减小噪音句子的权重；最后在真实的数据集上，通过计算准确率、召回率，绘出 PR 曲线证明该方法与现有的一些方法相比，取得了比较显著的进步。

关键词：深度学习；远程监督；GRU；注意力机制

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0197

Distant supervision relationship extraction based on GRU and attention mechanism

Huang Zhaowei, Chang Liang, Bin Chenzhong†, Sun Yanpeng, Sun Lei

(*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin Guangxi* 541004, *China*)

Abstract: With the development of deep learning, more and more deep learning models have been applied to the task of relation extraction, but traditional deep learning models cannot solve long distance dependence problems. At the same time, distant supervision will inevitably generate wrong labels. For these two problems, this work proposes a distant supervision relationship extraction method based on GRU (Gated Recurrent Unit) and the attention mechanism. First, the GRU neural network is adopted to extract text features and solve long-distance dependence problems. Second this work constructs a Sentence-Level Attention Mechanism on entity pairs to reduce the weight of noise sentences. Finally, based on the real data set, by calculating the accuracy rate and recall rate, the PR curve is drawn to prove the proposed method has achieved significant progress compared with some existing methods.

Key words: deep learning; distant supervision; GRU; attention mechanism

0　引言

关系抽取[1]不仅是信息抽取的任务之一，也是构建和补全知识图谱的关键所在，其研究的主要内容是从文本内容中挖掘出实体与实体之间的语义关系[2]，从纯文本生成关系数据的过程，是自然语言处理（NLP）中的关键任务。该任务可以描述为：给定一段文本 S，确定两个目标实体对 $<e_1,e_2>$ 的关系类别 r。

近年来，Freebase[3]、DBpedia[4]、YAGO[5]等知识库的建立，在 NLP 任务中得到广泛的应用，包括搜索、推荐、问答等。这些知识库都是由三元组事实构成，如(姚明，老婆，叶莉)。尽管现有的知识库中已经包含了大量的知识，但随着大数据时代的到来，这些知识仍无法满足人们的需要，迫切的需要从无结构化的文本中提取出结构化的数据。

本文提出了一种 GRU（gated recurrent unit）神经网络和注意力机制相结合的远程监督关系抽取方法（GRU_ATT）。采用远程监督的方法，避免了人工在构建数据集上花费的时间与精力，采用 GRU 模型克服了传统深度学习模型无法解决长距离依赖[6]的问题，同时在句子层面上引入注意力机制，有效控制了冗杂数据给实验结果带来的影响。最后，在真实的数据集上评估 GRU_ATT，实验结果表明，与现有方法相比，GRU_ATT 在关系提取方面取得了一些明显的进步。文章的主要贡献在于：a)使用一种在句子层面构建注意力机制的方法;b)提出一种将注

收稿日期：2018-03-19；**修回日期：**2018-04-28　**基金项目：**国家自然科学基金资助项目（U1501252，61572146）；广西创新驱动重大专项项目（AA17202024）；广西自然科学基金资助项目（2016GXNSFDA380006）；广西信息科学实验中心平台建设项目（PT1601）

作者简介：黄兆玮（1991-），男，四川成都人，硕士，主要研究方向为知识图谱与智能系统；常亮（1980-），男，贵州毕节人，教授，博士，主要研究方向为数据与知识工程，智能规划，形式化方法；宾辰忠（1979-），男（通信作者），湖南衡山人，讲师，博士，主要研究方向为数据挖掘与智能推荐（316413286@qq.com）；孙彦鹏（1993-），男，河北邢台人，硕士，主要研究方向为数据挖掘与推荐系统；孙磊（1992-），男，河南驻马店人，硕士，主要研究方向为位置感知与模式挖掘.

意力机制与 GRU 相结合的关系抽取方法。

1　相关工作

传统的关系分类的方法大多都是基于模式匹配，例如 Harabagiu 等人[7]提出一种通过结合词汇和语义关系来进行关系分类的方法；Kambhatla[8]提出一种基于逻辑回归的关系抽取方法。这些基于模式匹配方法都取得了较好的性能。但是不足之处在于：许多传统的 NLP 系统被用来提取高级特征，如词性标签、最短依赖路径和命名实体，从而导致计算成本的增加和额外的传播误差。另一个缺点是，由于不同的训练数据集的覆盖率较低，在通用性方面表现不佳。

大多数现有的受监督关系提取方法需要大量的标记关系训练数据，这样非常耗时费力。2009 年 Mintz 等人[9]首次提出了远程监督，他们通过对齐三元组知识库（KB）和文本，自动生成训练数据集。他们假设：如果两个实体在 KB 中有关系，那么包含这两个实体的所有句子将表达这种关系。例如，（姚明，老婆，叶莉）是 KB 中的三元组事实。远程监督将把包含这两个实体的所有句子视为该关系的正例。尽管远程监督是自动标记训练数据的有效策略，但它还是会带来一些错误的标签问题。例如，“姚明被邀参加腾讯体育名人赛，叶莉表示非常支持”“姚明与叶莉共同出现在上海浦东国际机场”，虽然这两个句子中都出现了“姚明”和“叶莉”，但却都没有表达出“老婆”的关系，在远程监督中仍将被视为一个正例。因此，Riedel 等人[10]采取多示例问题的处理办法，将一个关系所标记的所有句子看作是一个集合，并假设集合中至少有一个句子能表达出两个实体所对应的关系，从而有效的减少了冗杂数据对应远程监督的影响。文献[11]提出了基于概率图模型的多实例多标签模型，不仅对噪声训练数据进行建模，而且针对实体对和所属关系进行多分类建模，实验结果表明该模型在关系抽取效果上有了比较显著的提高。Liu 等人[12]提出一种基于卷积神经网络的弱监督关系抽取方法，将关系抽取技术运用到特定的领域中。他们构建了半自动化模式抽取系统 EARES，通过该系统生成训练语料，转换成向量特征矩阵，最后使用卷积神经网络（CNN）进行分类模型训练实现关系提取。Santos 等人[13]提出了一种在没有手工特征的情况下使用深度神经网络进行关系分类的方法，该方法是基于句子构建分类器，由于缺乏人工注释的训练数据，不能在大规模的知识库中应用。Zeng 等人[14]将多实例学习与神经网络模型相结合，建立了基于远程监督的关系提取器，虽然该方法在关系提取方面取得了显着的改进，但是效果还远远不能令人满意。

注意力机制和人类的选择性视觉注意力机制类似，核心目标也是从众多信息中，通过计算概率分布，选择出对当前任务目标更关键的信息。文献[15]在使用 RNN 模型做图像分类时，加入了注意力机制，之后 Bahdanau 等人[16]使用类似的注意力机制在机器翻译任务上，将翻译和对齐同时进行，这也是注意力机制首次应用到 NLP 领域中。Lin 等人[17]提出了一种基于远程监督关系提取的方法，使用 CNN 来提取句子特征，之后为解决错误的标签问题，构建句子层面的注意力机制。

2　关系抽取模型 GRU_ATT

模型总体结构如图 1 所示，首先得到包含同一对实体的所有句子集合 $S=\{s_1,s_2,s_3,...,s_n\}$，将 s_i 转化成向量表示 x_i，再使用 GRU 来提取句子的语义，得到句子的语义特征向量 X_i 之后，再通过注意力机制计算得到相应的权重 θ_i，最后将句子的语义特征向量点乘它们的权重 θ_i 再求和，得到集合 S 的向量表示 S_v。

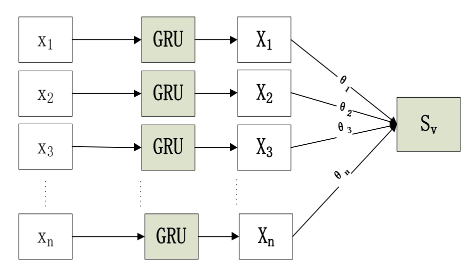


图 1　基于注意力机制和 GRU 神经网络的关系抽取模型

2.1　向量化

2.1.1 词向量化

给定一个由 t 个单词组成的句子 $s=\{w_1,w_2,w_3,...w_t\}$，运用 word2vec[18]将每一个单词 w_i 映射到一个低维实值向量空间中，将句子中的每个单词的向量拼接起来形成句子的向量。通过式（1）对句子进行词向量处理。

$$x_i = W^{word} \cdot V^i \quad (1)$$

其中：e_i 是词 w_i 的向量形式，$W^{word} \in R^{d^w|m|}$ 是句子的向量矩阵，其中 m 是一个固定尺寸的单词表，d^w 表示词向量的维数，V^i 是词 w_i 的 one-hot 表示。最终得到句子中每个单词的向量化表示：$V_S=\{x_1,x_2,x_3,...x_t\}$。

2.1.2 位置向量化

在关系抽取的任务中，往往靠近实体的单词更能突显出句子中两个实体之间的关系，因此为了能够更加准确地表达句子的含义，将句子中每一个单词到两个实体的距离拼接到该单词的词向量表示中。若在句子向量化中，词向量化的维度为 d^w，位置向量化的维度为 d^p，则句子向量的维度为

$$d^s = d^w + 2d^p \quad (2)$$

例如“Beijing is the capital of China”向量化如图 2 所示，在这个句子中“Beijing”和“China”作为两个实体，那么“is”到“Beijing”的距离的为-1，到“China”的距离为 4；“the”到“Beijing”的距离为-2，到“China”的距离为 3。

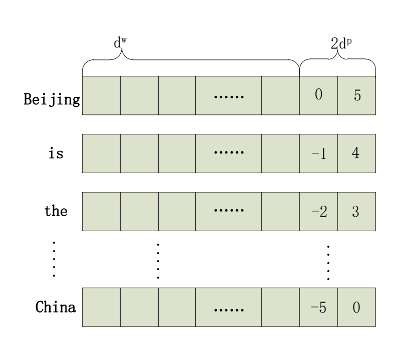


图 2　句子向量化

2.2　构建 GRU

LSTM（long short term memory）是一种改进的 RNN，可以学习长期依赖信息，由 Hochreiter & Schmidhuber 在 1997 年提出，最近又被 Alex Graves 进行了改良和推广。GRU[19]是 LSTM 的一种变体，保持了 LSTM 的效果，同时又使结构更加简单，减少了训练参数，提高了模型训练的速率。

LSTM 是一种复杂的网络结构，包含三个门计算即忘记门、输入门和输出门。三个门计算用来增加或者减少细胞状态中的信息，GRU 将 LSTM 中的忘记门和输入门合成为一个单一的更新门，同时还混合了细胞状态和隐藏状态，更新门用于控制前一时刻的信息是否被带入到当前细胞状态中，GRU 结构如图 3 所示。

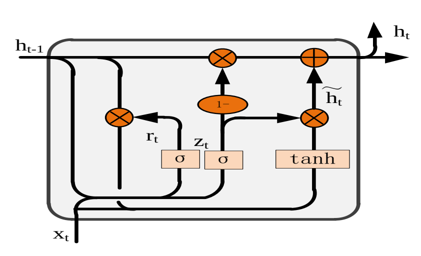


图 3　GRU 模型

将上一节中得到的句子的向量化表示作为该步输入，以第 i 个单词为例，说明第 i 个单词在 GRU 单元各个状态的特征值：

$$z_t = rec(W_{hz}h_{t-1} + W_{xz}x_t + b_z) \tag{3}$$

$$r_t = rec(W_{hr}h_{t-1} + W_{xr}x_t + b_r) \tag{4}$$

$$\tilde{h}_t = \tanh(W_{rh}r_t h_{t-1} + W_{wx}x_t + b_h) \tag{5}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{6}$$

其中：rec 激活函数选用 relu 函数，X_t 表示当前时刻的输入，h 表示记忆体，W 表示权值矩阵，b 表示偏置量。

2.3　引入注意力机制

假设包含实体对 $<e_1, e_2>$ 所有句子的集合 $S = \{s_1, s_2, s_3, ..., s_n\}$，本小节中为了充分利用集合中每一个句子的信息，引入注意力机制来计算注意力概率，从而体现集合中某一个句子对于集合的重要程度。假设现在判断实体对 $<e_1, e_2>$ 是否具有关系矢量 $\boldsymbol{r}$，那么集合 S 中的每一个句子 s_i 都包含了该句子是否表达了 $\boldsymbol{r}$ 关系的信息，首先，将集合 S 转化为向量的形式 S_v，计算公式如下：

$$S_v = \sum_i \beta_i x_i \tag{7}$$

β_i 表示句子 x_i 的权重。为了验证注意力机制的引入给实验结果带来的影响，用两种方式定义 β_i：第一种，令 $\beta_i = \frac{1}{n}$，也就是说，认为集合中所有的句子对于表达关系 $\boldsymbol{r}$ 都同等重要。显然，这样是不合理的，因为集合中总会有句子不表达关系 $\boldsymbol{r}$，这种冗杂数据就会给实验的结果产生不好的影响。第二种，为了避免这种情况的发生，减小这种噪音数据对实验结果的影响，准确给出集合中每个句子的权重，首先定义一个关于 s_i 和 $\boldsymbol{r}$ 的函数：

$$scores(i) = s_i A r \tag{8}$$

该函数描述集合中某一个句子 S_i 与预测关系 $\boldsymbol{r}$ 之间的匹配程度,将该函数的取值范围定义在 $[0,1]$，0 表示某个句子 s_i 完全不可能表达 $\boldsymbol{r}$ 关系；相反，1 表示某个句子一定会表达 $\boldsymbol{r}$ 关系。$\boldsymbol{A}$ 为一个对角矩阵。接下来将式（8）通过 softmax function 就得到第 i 个句子在集合中的权重：

$$\beta_i = \frac{\exp(scores(i))}{\sum_k \exp(scores(k))} \tag{9}$$

接着将式(9)代入式(7),就得到了集合 $\boldsymbol{S}$ 的向量形式:

$$S_v = \sum_i \frac{\exp(scores(i))}{\sum_k \exp(scores(k))} x_i \tag{10}$$

再得到 S_v 后，接下来定义一个线性函数就可以计算每一个可能的关系 $\boldsymbol{r}$ 的得分：

$$y = MS_v + b \tag{11}$$

M 是一个关系矩阵，b 是一个偏置量。

2.4　模型训练与优化求解

采用最小化负对数似然函数来进行模型训练，接着将式（11）通过 softmax 层定义条件概率如下：

$$P(r \mid S, \alpha) = \frac{\exp(y_r)}{\sum_{j=1}^{n_r} \exp(y_j)} \tag{12}$$

n_r 表示所有可能的关系。接下来采用随机梯度下降算法来最小化负对数似然函数，定义优化目标函数 $J(\alpha)$ 计算如下：

$$J(\alpha) = -\sum_{i=1}^{D} \log p(r_i \mid S_i, \alpha) = -\sum_{i=1}^{D} \log\left(\frac{\exp(y_{r_i})}{\sum_{j=1}^{n_r} \exp(y_j)}\right) \tag{13}$$

其参数 $\alpha(M, b, A)$ 表示模型训练参数，初始值随机给定。D 表示训练样本数。

3　相关工作实验结果

3.1　实验数据与实验环境

实验环境：操作系统 windows7，64 位；处理器 Intel(R) Core™ i5-4690；内存大小为 8 GB；编程平台 Pycharm，Python2.7 版本。

实验数据：本次实验采用的数据集，是通过将 Freebase 中的实体对与纽约时报语料库（NYT）对齐而生成的。该数据集[10]是 Riedel 在 2010 年首次开发使用的，之后也被 Raphael Hoffmann、Mihai Surdeanu、Lin 等人使用，该数据集使用斯坦福大学的命名实体识别工具对纽约时报语料库进行实体标注，然后再与 Freebase 中的实体进行匹配，生成数据集格式如表 1 所示。第一列中的两行分别表示第一个实体和第二个实体的 id，第二列中的两行分别表示两个实体，第三列表示关系，第四列表示句子。假设（Hunan，contains，changsha）就是表示 Freebase 中的一条三元组数据：“one reason is that hunan's fast-growing provincial capital changsha is beginning to siphon some workers”就是纽约时报语料库的一个句子，将 Hunan 和 changsha 与句子中相同实体对应起来就行了一条数据。

该数据集包含特殊关系 NA（NA:句子中的两个实体没有任何关系）在内，一共 53 种关系、39528 个实体。训练集包含 522611 个句子，281270 个实体对，18252 个关系事实；测试集包含 172448 个句子，96678 个实体对，1950 个关系事实。

表 1 数据集格式

实体 Aid/ 实体 Bid	实体 A/ 实体 B	关系	句子
m.01669t/ m.01cw6l	Hunan/ changsha	../contains	one reason is that hunan's fast-growing provincial capital changsha is beginning to siphon some workers
m.0dlw0/ m.03rjj	Umbria/ italy	../country	it echoed the insistence of his grandfather an immigrant from umbria in italy that nary a word of italian be heard in the family's new home in nice
m.045c7b/m.01cw6	Google/ youtube	NA	Since it bought youtbe last october google has been chasing deals that would give it the right to put mainstream video programming on the site .

3.2 评测指标

本文通过采集准确率（precision）、召回率（recall）绘制 PR 曲线，作为评价指标。准确率、召回率计算公式如下：

$$precision=\frac{right_num}{out} \tag{14}$$

$$recall=\frac{right_num}{all} \tag{15}$$

right_num 表示模型预测正确的数据条数；*out* 表示一共预测的数据条数；*all* 表示测试集总共的数据条数。在准确率相同的情况下召回率越高越好，在召回率相同的情况下准确率越高越好，因此 PR 曲线越在右上方指标越好。

3.3 参数设置

采用 Softmax 作为分类器。采用 L2 正则化方法对网络参数进行约束，训练过程引入了采用 dropout 策略，采用批量的 Adadelta 优化方法用于模型训练，具体参数设置如表 2 所示。

表 2 参数设置

Vocab_size	16691
Steps_num	70
Classes_num	12
Gru_size	230
Layers_num	1
Epochs_num	10
Dropout_rate	0.5

3.4 实验结果与分析

3.4.1 与现有方法比较

首先将 GRU_ATT 与传统基于特征的关系抽取方法相比较得到图 4 所示实验结果。红色曲线代表 GRU_ATT 的实验结果；绿色曲线是 Mintz[9]，2009 年提出的传统基于远程监督的关系抽取模型。黑色曲线是 MultiR[20]，2011 年 Hoffmann 提出的处理重复关系的多实例学习模型。在整个召回范围内，GRU_ATT 都好于 Mintz 与 MultiR 两种基于特征的方法。Mintz 和 MultiR 在召回大于 0.2 左右的时候开始迅速下降，而 GRU_ATT 在整个过程中都相对稳定，这表明通过人工设计的特征不能够准确的表达出句子的语义。

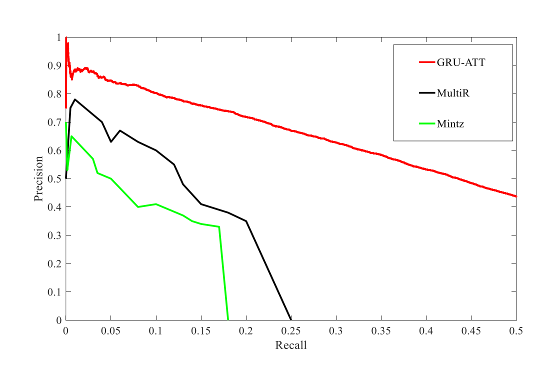


图 4 GRU_ATT 与传统基于特征的方法比较

为了显示 GRU 模型的优势，将该方法与 2015 年 Lin 提出了一种基于远程监督关系提取的方法[17]CNN_ATT 相比较，他们通过 CNN 来提取句子特征，之后为解决错误的标签问题，构建句子层面的注意力机制。GRU_ATT 与 CNN_ATT 对比实验结果如图 5 所示。

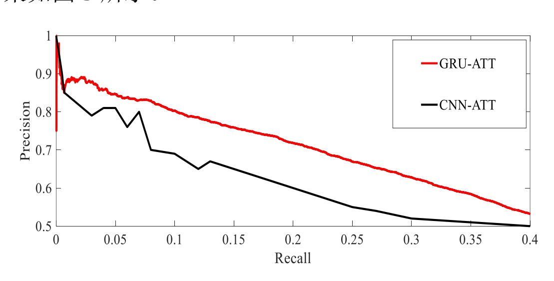


图 5 GRU_ATT 与 CNN_ATT 对比实验结果

黑色线条是 CNN_ATT，红色线条是 GRU_ATT。在该任务

中，CNN 只能抽取位置不变的特征，不能学习时序序列。而 GRU 因为有了记忆模块，能够充分利用整个句子的序列信息，包括各个词之前的关联信息，更适合自然语言处理的任务。因此，由实验结果可以看出，GRU_ATT 要好于 CNN_ATT。

3.4.2 注意力机制对实验结果产生的影响

该实验部分是为了验证注意力机制的引入给实验结果带来的影响。设定三个模型，第一个模型只采用 GRU，不引入注意力机制；第二个模型采用 GRU_AVE，也就是 $\beta=\frac{1}{n}$，将同对实体对的所有句子的权重都看做一样；第三个模型就是 GRU_ATT，对比实验结果如图 6 所示。

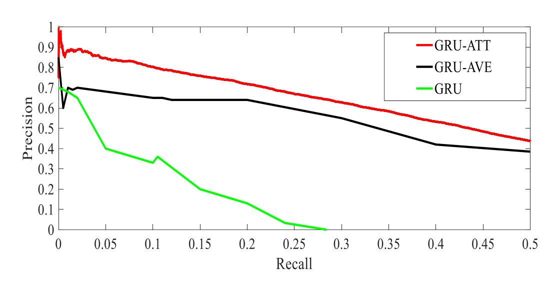


图 6　注意力机制对实验结果的影响

红色线条代表 GRU_ATT；黑色线条代表 GRU_AVE；绿色线条 GRU。GRU_AVE 引入注意力机制，考虑了句子表达的意义，从而减小了冗杂数据对于实验结果的影响，而 GRU 没有引入注意力机制，完全不考虑噪声数据的影响，因此 GRU_AVE 的实验结果要好于 GRU。但是 GRU_AVE 将所有的句子都看做是平等的，因此也会带入一些冗杂数据，那些表达错误关系的句子就对实验结果产生了负面的影响。GRU_ATT 在整个召回范围实现了最高的精度，这表明注意力机制的引入可以有效的减小远程监督中冗杂数据对于实验结果的影响。

4　结束语

本文提出了一种新的神经网络模型 GRU_ATT，用于文本关系提取。该模型克服了传统深度学习模型无法解决长距离依赖问题，大大减小了人工对数据标记所耗费的时间和精力，减小了冗杂数据对于实验结果的影响。在公共语料库上的实验结果，PR 曲线也有所提高。尽管由于注意力机制的引入，是减小了冗杂数据对于实验结果的影响，但该问题并没有完全解决。下一步工作，将着重考虑如何在开放领域自动发现关系，进行关系抽取。

参考文献：

[1] 赵妍妍，秦兵，车万翔，等．中文事件抽取技术研究 [J]. 中文信息学报, 2008, 22 (1): 3-8. (Zhao Yanyan, Qin Bing, Chen Wanxiang, Liu Ting, *et al.* Research on Chinese Event Extraction [J]. Journal of Chinese Information Processing, 2008, 22 (1): 3-8.)

[2] 陈宇，郑德权，赵铁军．基于 Deep Belief Nets 的中文名实体关系抽取 [J]. 软件学报，2012，23 (10): 2572-2585. (Chen Yu, Zheng Dequan+, Zhao Tiejun. Chinese Relation Extraction Based on Deep Belief Nets [J]. Journal of Software, 2012, 23 (10): 2572-2585.)

[3] Huang Xun, You Hongliang, Yu Yang. A Review of Relation Extraction [J]. NewTechnology of Library and Information Service, 2013 (11): 30-39

[4] Kurt B, Colin E, Praveen P, *et al.* Freebase: a collaboratively created graph database for structuring human knowledge [C]// Proc of KDD. 2008: 1247–1250.

[5] Lehmann J. DBpedia: a nucleus for a web of open data [C]// Proc of Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference. 2007: 11-15.

[6] Hochreiter S, Schmidhuber J. Long short-term memory. [J]. Neural Computation, 1997, 9 (8): 1735.

[7] Rink B, Harabagiu S. UTD: classifying semantic relations by combining lexical and semantic resources [C]// Proc of International Workshop on Semantic Evaluation. Association for Computational Linguistics. 2010: 256-259.

[8] Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]// Interactive Poster and Demonstration Sessions. Stroudsburg: Association for Computational Linguistics, 2004: 22.

[9] Mintz M, Steven B, Rion S, *et al.* Distant supervision for relation extraction without labeled data [C]// Proc of Joint Conference of the Meeting of the ACL and the, International Joint Conference on Natural Language Processing of the Afnlp: Volume. Stroudsburg: Association for Computational Linguistics, 2009: 1003-1011.

[10] Riedel S, Yao Limin, Mccallum A. Modeling relations and their mentions without labeled text [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2010: 148-163.

[11] Surdeanu M, Tibshirani J, Nallapati R, *et al.* Multi-instance multi-label learning for relation extraction [C]// Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2010: 455-465.

[12] 刘凯，符海东，邹玉薇，等．基于卷积神经网络的中文医疗弱监督关系抽取 [J]. 计算机科学，2017, 44 (10): 249-253. (Liu Kai, Fu Haidong, Zou Yuwei, Gu Jinguang, *et al.* Chinese medical weak supervised relation extraction based on convolution neural network [J]. Computer Science, 2017, 44 (10): 249-253.)

[13] Santos C N D, Gattit M. Deep convolutional neural networks for sentiment analysis of short texts [C]// Proc of International Conference on Computational Linguistics. 2014

[14] Zeng Daojian, Liu Kang, Chen Yubo, *et al.* Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[15] Mnih V, Heess N, Graves A, *et al.* Recurrent models of visual attention

[EB/OL]. (2014-12-03) . https://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention. pdf.

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014.

[17] Lin Yankai, Shen Shiqi, Liu Zhiyuan, *et al*. Neural relation extraction with selective attention over instances [C]// Proc of Meeting of the Association for Computational Linguistics. 2016: 2124-2133.

[18] Mikolov T, Sutskever I, Chen Kai, *et al*. Distributed representations of words and phrases and their compositionality [EB/OL]. (2014-11-25) . https://papers.nips.cc/paper/5021-distributed-representations-of-words-and -phrases-and-their-compositionality. pdf.

[19] Dey R, Salemt F M. Gate-variants of Gated Recurrent Unit (GRU) neural networks [C]// Proc of IEEE nternational Midwest Symposium on Circuits and Systems. 2017: 1597-1600.

[20] Hoffmann R, Zhang Congle, Ling Xiao, *et al*. Knowledge-based weak supervision for information extraction of overlapping relations [C]// Proc of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 541-550.